Adaptive estimation of a conditional density

Ann-Kathrin Bott* and Michael Kohler.
Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: abott@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de

December 16, 2014

Abstract
In this paper we estimate a conditional density by a conditional kernel density estimate. The error of the estimate is measured by the L_1–error. Based on the combinatorial method of Devroye and Lugosi (1996) we propose a new method for choosing the bandwidths adaptively and derive a theoretical result about the quality of this method. Moreover we illustrate the performance of the estimate for finite sample size by using simulated data.

AMS classification: Primary 62G07; secondary 62G20.

Key words and phrases: Conditional density estimation, L_1–error, bandwidth selection.

1 Introduction

One major problem in statistics is the estimation of a distribution from a given sample. Let Z be a $\mathbb{R}^d$-valued random variable with distribution μ and let $Z_1, \ldots, Z_n$ be an independent sample of Z. According to the theorem of Glivenko-Cantelli the empirical distribution function

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(Z_i) \quad (A \in \mathcal{B}_d)$$

is a consistent estimate of the probabilities of all intervals. However, if we are interested in an estimation of general sets, we want to construct estimates $\hat{\mu}_n$ such that the total variation error

$$\sup_{B \in \mathcal{B}_d} |\hat{\mu}_n(B) - \mu(B)| \tag{1}$$

(where $\mathcal{B}_d$ denotes the Borel sigma-field) converges to zero almost surely. It is well known, that there does not exist any estimate $\hat{\mu}_n$ such that

$$\sup_{B \in \mathcal{B}_d} |\hat{\mu}_n(B) - \mu(B)| \to 0 \quad \text{a.s.}$$

for all distributions (cf., Devroye and Györfi (1990)). But if μ has a density f with respect to the Lebesgue-Borel measure, then we can construct universally L_1–consistent

*Corresponding author. Tel: +49-6151-16-75878

density estimates f_n, i.e., estimates f_n satisfying

$$\int |f_n(x) - f(x)|\, dx \to 0 \quad \text{a.s.}$$

for all densities (cf., e.g., Devroye (1983)). E. g., the Rosenblatt-Parzen kernel density estimate defined by

$$f_n(x) = \frac{1}{n \cdot H_n} \cdot \sum_{k=1}^{n} K\left(\frac{\|x - Z_k\|}{H_n}\right)$$

with density K and bandwidth $H_n > 0$ (cf., e.g., Rosenblatt (1956) and Parzen (1962)) has this property whenever the bandwidth is chosen such that

$$H_n \to 0 \quad (n \to \infty) \quad \text{and} \quad n \cdot H_n^d \to \infty \quad (n \to \infty)$$

(cf., e.g., Mnatsakanov and Khmaladze (1981) and Devroye (1983); general results in density estimation can be also found in the books of Devroye and Györfi (1985), Devroye (1987) and Devroye and Lugosi (2001)). If f_n is a density, then the Lemma of Scheffé states that

$$\sup_{B \in \mathcal{B}_d} |\hat{\mu}_n(B) - \mu(B)| = \frac{1}{2} \int |f_n(x) - f(x)|\, dx,$$

where

$$\hat{\mu}_n(B) = \int_B f_n(x)\, dx$$

is the corresponding distribution estimate, hence, the corresponding distribution estimate enables a consistent estimation of the probability of all sets.

Like the kernel density estimate most estimates depend on parameters. For instance, the histogram estimate depends on the partition of $\mathbb{R}^d$. Considering a finite sample the parameter choice is of great interest. Let $Z_1, \ldots, Z_n$ be an independent sample of an $\mathbb{R}^d$–valued random variable Z with density f. Moreover we assume that a class of density estimates $(f_{n,\theta})_{\theta \in \Theta}$ is given. Now we want to choose a parameter $\hat{\theta} \in \Theta$ such that

$$\int |f(x) - f_{n,\hat{\theta}}(x)|\, dx \approx \inf_{\theta \in \Theta} \int |f(x) - f_{n,\theta}(x)|\, dx.$$

Due to the fact that f is unknown, the L_1–error cannot be determined. This raises the question how to select parameters in order to minimize the L_1–error. Typically this question is considered in the literature in connection with the L_2–error, see, e.g., Rudemo (1982), Hall (1983), Bowman (1984), Stone (1984), Hall et al. (1991) and the literature cited therein.

But much less is known concerning adaptation result in connection with the L_1–error. In this respect Devroye and Lugosi (1996) intoduced the so called combinatorial method to choose the parameters of a density estimate in dependence of the given sample. At first, the sample is splitted into testing data $Z_1, \ldots, Z_m$ and learning data $Z_{m+1}, \ldots, Z_n$ for $0 < m \leq \lfloor n/2 \rfloor$. The learning data is used to define the estimate which is denoted by

$f_{n-m,\theta}(\cdot) = f_{n-m,\theta}(\cdot\,, Z_{m+1}, \ldots, Z_n)$. The empirical distribution function of the testing data is defined as

$$\mu_m(A) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_A(Z_i) \quad (A \in \mathcal{B}_d).$$

The combinatorial method chooses the parameter $\hat{\theta} \in \Theta$ for which the expression

$$\Delta_\theta = \sup_{A \in \mathcal{A}} \left| \int_A f_{n-m,\theta}(x)\, dx - \mu_m(A) \right| \tag{2}$$

is minimal, where $\mathcal{A}$ denotes the Yatracos class of subsets of $\mathbb{R}^d$, given by

$$\mathcal{A} = \left\{ \left\{ x \in \mathbb{R}^d : \, f_{n-m,\theta_1}(x) > f_{n-m,\theta_2}(x) \right\} \, : \theta_1, \theta_2 \in \Theta \right\}.$$

Devroye and Lugosi (1996) showed that the L_1–error of the resulting estimate $f_{n-m,\hat{\theta}}$ is linked to the L_1–error with the optimal parameter choice. If $\int f_{n-m,\theta}(x)\, dx = 1$ for all $\theta \in \Theta$, it holds

$$\int |f(x) - f_{n-m,\hat{\theta}}(x)|\, dx \leq 3 \cdot \inf_{\theta \in \Theta} \int |f(x) - f_{n-m,\theta}(x)|\, dx + 4\Delta + \frac{3}{n}, \tag{3}$$

where

$$\Delta = \sup_{A \in \mathcal{A}} \left| \int_A f(x)\, dx - \mu_m(A) \right|.$$

If the condition $\int f_{n-m,\theta}(x)\, dx = 1$ is not fulfilled for all $\theta \in \Theta$, the statement also holds but with factor "5" instead of "3". In addition, Devroye and Lugosi (1997a) derived upper bounds for $\mathbf{E}\{\Delta\}$ by combinatorial tools. For a suitable choice of m the last two summands are asymptotically neglectable. Hence, the L_1–error can be bounded by a multiple of the L_1–error of the estimate with the optimal bandwidth. In Chapter 11 of Devroye and Lugosi (2001) concrete results for the classes of kernel density estimates are summarised. A comparison to other methods and simulation results are given in Devroye and Lugosi (1997b).

In this paper we deal with conditional density estimation. Here, one is interested in the conditional density of a random variable Y given a random vector X. This problem can be seen as generalization of regression. One is interested in the full density rather than in the expected value. In conditional density estimation it is usually assumed that a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of an $\mathbb{R}^d \times \mathbb{R}$–valued random vector (X, Y) is available. Already in Rosenblatt (1969) the kernel estimate of a conditional density was introduced. But it first received serious attention in Fan et al. (1996) and Hyndman et al. (1996). This estimator is motivated by the definition of a conditional density. Let $f_{(X,Y)}(x, y)$ be the joint density of (X, Y) and $f_X(x)$ the marginal density of X. Then the conditional density $f_{Y|X}(y, x)$ of Y given X is given by

$$f_{Y|X}(x, y) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}.$$

Replacing the joint and marginal density by density estimates we obtain an estimator of the conditional density. The marginal density of X can by estimated by the Rosenblatt-Parzen kernel density estimate given by

$$f_n(x) = \frac{1}{n \cdot H_n^d} \cdot \sum_{k=1}^{n} K\left(\frac{\|x - X_k\|}{H_n}\right) \tag{4}$$

with density K and bandwidth $H_n > 0$. Using the product kernel estimator (cf., e.g., Rosenblatt (1969), Scott (1992) and Hyndman et al. (1996)) we can estimate the joint density by

$$\hat{f}_{(X,Y)}(x,y) = \frac{1}{n \cdot H_n^d \cdot h_n} \sum_{i=1}^{n} K\left(\frac{\|x - X_i\|}{H_n}\right) \cdot K\left(\frac{|y - Y_i|}{h_n}\right)$$

where $K : \mathbb{R} \to \mathbb{R}_+$ is a density and h_n, $H_n > 0$ are bandwidths. Hence, we can estimate the conditional density by

$$\hat{f}_{Y|X}(y,x) = \frac{\sum_{i=1}^{n} K\left(\frac{\|x-X_i\|}{H_n}\right) \cdot K\left(\frac{|y-Y_i|}{h_n}\right)}{h_n \sum_{j=1}^{n} K\left(\frac{\|x-X_j\|}{H_n}\right)}. \tag{5}$$

This conditional density estimation problem can also be seen as a nonparametric regression problem. It is well known that

$$\mathbf{E}\left\{\frac{1}{h_n} \cdot K\left(\frac{y - Y}{h_n}\right) \middle| X = x\right\} \to f_{Y|X}(y,x) \quad (n \to \infty)$$

for Lebesgue almost all y and $\mathbf{P}_X$-almost all x (cf., e.g., Fan et al. (1996)). Thus, the estimator (5) can be seen as a kernel regression estimate (cf., e.g., Chapter 5 in Györfi et al. (2002)) applied to

$$\left(X_1, \frac{1}{h_n} \cdot K\left(\frac{y - Y_1}{h_n}\right)\right), \ldots, \left(X_n, \frac{1}{h_n} \cdot K\left(\frac{y - Y_n}{h_n}\right)\right),$$

cf., e.g., Fan and Yim (2004) and Gooijer (2003). Also other regression estimates can be applied to this setting, for instance, Györfi and Kohler (2007) considered a partitioning estimate.

The question now arises how to adaptively select the bandwidths. Literature only deals with methods concerning the L_2–error. Various attempts start with choosing the bandwidth h_n by referencing rules and afterwards select H_n by known methods for kernel regression estimate. Fan et al. (1996) choose h_n by the normal referencing rule of Silverman (1986) and H_n by the residual squares criterion (Fan and Gijbles (1996)). Also Bashtannyk and Hyndman (2001) and Hyndman and Yao (2002) first apply bandwidth rules based on a reference distribution to determine one of the bandwidths and then apply regression based bandwidth selectors to determine the second one. These methods

use strong assumptions on the distributions and therefore work only well in a limited number of cases. Hall et al. (1999) proposed a bootstrap method, that works well for polynomial regression models. Also Bashtannyk and Hyndman (2001) considered this approach and extended the method. Fan and Yim (2004) proposed a method without restrictive assumptions. They choose the bandwidth by cross-validation. While the first mentioned ad-hoc methods can be efficiently calculated, they perform poorly on finite samples for most distributions. On the other hand the bootstrap method and cross-validation method are time-consuming but more reliable. Holmes et al. (2010) try to balance between both aspects and proposed a likelihood cross-validation method.

In this paper we derive and analyze a data dependent method to choose the bandwidths $h_n, H_n > 0$ of a conditional kernel estimate without any assumptions on the distribution of (X, Y). This method is motivated by the above mentioned combinatorial method of Devroye and Lugosi (1996). Since we do not estimate one single density, we transform Δ_θ such that the resulting adaptive estimate is an appropriate estimate of $f(\cdot, x)$ for $\mathbf{P}_X$-almost all $x \in \mathbb{R}^d$. The main difficulty here is that we estimate simultaneously $f(\cdot, X_i)$ for $i \in \{1, \dots, n\}$, where for each i we have available only a sample of size one which we cannot split into learning and testing data.

Since we are interested in an estimation of the conditional distribution of Y given X, we measure the quality of the adaptive estimate by the L_1–error. More precisely, we consider the average L_1–error

$$\int \int |f_n(y, x) - f(y, x)| \, dy \, \mathbf{P}_X(dx),$$

and we show that the expected average L_1–error of our newly proposed adaptive estimate is (up to a term of order $\sqrt{\log(n)}/\sqrt{n}$) less than or equal to five times the expected L_1–error which we would get if we would be able to choose the bandwidth in an optimal way (which is never possible in an application). The proofs are based on a generalization of results from empirical process theory to a setting where the data is independent but not identically distributed, which requires various non-trivial modifications of known techniques.

Throughout the paper the following notation is used: The sets of natural numbers, integers, real numbers and positive real numbers including zero are denoted by $\mathbb{N}, \mathbb{Z}, \mathbb{R}$ and $\mathbb{R}_+$, respectively. $\mathcal{B}_d$ denotes the set of all Borel sets in $\mathbb{R}^d$ and $\mathbb{1}_B$ denotes the indicator function of the set B. $\|x\|$ is the Euclidean norm of a vector $x \in \mathbb{R}^d$. For a real number z we denote by $\lfloor z \rfloor$ and $\lceil z \rceil$ the largest integer less than or equal to z and the smallest integer larger than or equal to z, respectively.

The outline of this paper is as follows: The main results are presented in Section 2 and proven in Section 4. Section 3 illustrates the finite sample size behavior of our estimate by applying it to simulated data.

2 Main results

We assume that an independent and identically distributed sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of an $\mathbb{R}^d \times \mathbb{R}$–valued random vector (X, Y) is available. We select simultaneously the

bandwidths $h_n, H_n > 0$ of our estimate

$$f_n(y,x) = \frac{\sum_{i=1}^n K\left(\frac{\|x-X_i\|}{H_n}\right) K\left(\frac{|y-Y_i|}{h_n}\right)}{h_n \sum_{j=1}^n K\left(\frac{\|x-X_j\|}{H_n}\right)}$$

where $K(x) = \frac{1}{2} \cdot \mathbb{1}_{[-1,1]}(x)$ is the naive kernel. At first we choose a parameter set

$$\mathcal{P}_n \subseteq \left\{(h,H) \in \mathbb{R}^2 \middle|\, h \in [1/n,\, n],\, H > 0\right\}.$$

Now we split the data samples into two halves. The second half of the data $(X_{\lfloor n/2 \rfloor+1}, Y_{\lfloor n/2 \rfloor+1})$, $\ldots, (X_n, Y_n)$ is the so called learning data and is used to define our estimate:

$$\hat{f}_\theta(y,x) = \frac{\sum_{i=\lfloor n/2 \rfloor+1}^n K\left(\frac{\|x-X_i\|}{H}\right) K\left(\frac{|y-Y_i|}{h}\right)}{h_n \sum_{j=\lfloor n/2 \rfloor+1}^n K\left(\frac{\|x-X_j\|}{H}\right)}$$

with $\theta = (h,H)$. On the basis of the first half of the data (testing data) we evaluate our estimator and choose the parameters. Our goal is to select $\hat{\theta} \in \mathcal{P}_n$ such that the average L_1–error of the corresponding estimate $\hat{f}_{\hat{\theta}}$ is small. We select $\hat{\theta} = (\hat{h}, \hat{H})$ through minimizing

$$\Delta_\theta = \sup_{\theta_1,\theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int_{A_i(\theta_1,\theta_2)} \hat{f}_\theta(y, X_i)\, dy - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{1}_{A_i(\theta_1,\theta_2)}(Y_i) \right|,$$

where

$$A_i(\theta_1,\theta_2) = \left\{ y \in \mathbb{R} \,:\, \hat{f}_{\theta_1}(y, X_i) > \hat{f}_{\theta_2}(y, X_i) \right\}.$$

If the minimum does not exist, we choose $\hat{\theta} = (\hat{h}, \hat{H}) \in \mathcal{P}_n$ such that

$$\Delta_{\hat{\theta}} < \inf_{\theta \in \mathcal{P}_n} \Delta_\theta + \frac{1}{n}.$$

Δ_θ is motivated by (2). But here we consider the arithmetic mean of estimated L_1–errors. And since we estimate a whole class of densities, we need to regard a whole class of Yatracos sets, which are linked by the parameters $\theta_1, \theta_2 \in \mathcal{P}_n$.

The following theorem bounds the expected average L_1–error of this estimate by that of the estimate with optimal parameter choice.

Theorem 1 *Let $\hat{f}_{\hat{\theta}}$ be the above introduced estimate. It holds for all $n > 1$*

$$\begin{aligned}
&\mathbf{E}\left\{ \int \int |\hat{f}_{\hat{\theta}}(y,x) - f(y,x)|\, dy\, \mathbf{P}_X(dx) \right\} \\
\leq & 5 \cdot \inf_{\theta \in \mathcal{P}_n} \mathbf{E}\left\{ \int \int |\hat{f}_\theta(y,x) - f(y,x)|\, dy\, \mathbf{P}_X(dx) \right\} + \frac{2}{n} + 116 \sqrt{\frac{\log n}{\lfloor n/2 \rfloor}} + \frac{306}{\sqrt{\lfloor n/2 \rfloor \cdot \log n}}.
\end{aligned}$$

Remark 1. This theorem states that the expected average L_1–error of the proposed estimate lies close to five times the least possible. Here we have a factor of "5" instead of "3", since our estimate $\hat{f}_{\hat{\theta}}$ is (possibly) no density for all $x \in \mathbb{R}^d$. In case that for some $x \in \mathbb{R}^d$

$$\sum_{i=\lfloor n/2 \rfloor+1}^{n} K\left(\frac{\|x - X_j\|}{\hat{H}}\right) = 0,$$

it holds $\hat{f}_{\hat{\theta}}(y, x) = 0$ for all $y \in \mathbb{R}$.

Remark 2. Due to the splitting of the sample we compare the quality of our estimate to that of an estimate using also only half of the data. It is an open problem to show that

$$\inf_{\theta \in \mathcal{P}_n} \mathbf{E}\left\{\int\int |\hat{f}_\theta(y, x) - f(y, x)|\, dy\, \mathbf{P}_X(dx)\right\}$$

with $\hat{f}_\theta$ using half of the data is not much larger than with $\hat{f}_\theta$ using all of the data. Devroye and Lugosi (1997b) addressed this problem in case of density estimation (c.f., Devroye and Lugosi (1997b) and Theorem 10.3 in Devroye and Lugosi (2001)) .

Remark 3. By Theorem 1 a non-asymptotic upper bound of the expected average L_1–error is given. As we did not attempt to minimize the constants, the constants of the last two summands could potentially be much smaller.

3 Simulations

In this section we illustrate the performance of our estimator for finite sample size and finite parameter sets considering three examples. We compare the results to those of a conditional kernel estimate with cross-validated bandwidths like in Fan and Yim (2004). We evaluate the performance of both selection rules by the average L_1–error. The proposed estimate splits the data into learning and testing data. In Section 2 we assumed that $N = \lfloor n/2 \rfloor$ points were used to test the estimate and $n - N = \lceil n/2 \rceil$ data points to construct the estimate (new1). In addition we consider the proposed estimate with $N = \lfloor n/4 \rfloor$ testing data points and $n - N = \lceil 3n/4 \rceil$ learning data points (new2). To get an impression how small the average L_1–error could be under these circumstances, we compare our results to the estimate with n data points and the optimal parameter choice out of $\mathbf{h} \times \mathbf{H}$. In applications the underlying distribution is unknown and thus, this estimator is not applicable. In the implementation we approximate all integrals by Rieman sums. For each example an appropriate 10×10 grid $\mathbf{h} \times \mathbf{H}$ of bandwidths is considered and every method chooses the bandwidths out of this set.

In the first example 500 independent copies of (X, Y) are sampled, where X is uniformly distributed on $[0, 2]$ and Y is exponentially distributed with a rate depending on the covariate. More precise,

$$Y \sim Exp(\lambda) \quad \text{with } \lambda = 0.25 + X \text{ and } X \sim \mathcal{U}[0, 2].$$

The bandwidths are selected out of finite parameter sets

$$h \in \mathbf{h} = \{\, 0.1,\, 0.2,\, 0.3,\, 0.4,\, 0.5,\, 0.6,\, 0.7,\, 0.8,\, 0.9,\, 1.0\}$$

	optimal	**new1**	**new2**	**CV**
mean (sd) L_1–error	0.202 (0.019)	0.278 (0.041)	0.254 (0.038)	0.345 (0.089)
mean (sd) H	0.664 (0.130)	0.646 (0.300)	0.718 (0.347)	1.590 (0.044)
mean (sd) h	0.205 (0.036)	0.171 (0.078)	0.151 (0.080)	0.503 (0.376)

Table 1: Results of the first example.

$$H \in \mathbf{H} = \{0.2,\ 0.4,\ 0.6,\ 0.8,\ 1.0,\ 1.2,\ 1.4,\ 1.6,\ 1.8,\ 2.0\}.$$

To avoid edge effects we are only interested in estimating the conditional densities $f_1 : \mathbb{R} \times [0.5,\ 1.5] \to \mathbb{R}_+$, even though X is uniformly distributed on $[0, 2]$. Since the results of our simulation depend on randomly occuring data points, we repeat the whole procedure 100 times. The boxplots in Figure 1 report the average L_1–errors for all four estimates. Mean and standard deviation (sd) of the average L_1–errors as well as mean and standard deviation of the chosen bandwidths are given in Table 1. Here both proposed estimates outperform considerably the crossvalidated estimate. The second version of our estimate (new2) achieves even better results than the first version (new1). This results presumably from the higher amount of data that is used for the second estimate (new2). Even though the proposed estimates use less data than the crossvalidated estimate, the mean bandwidths are smaller.

Secondly, we let Y be normally distributed with mean corresponding to the covariate X and variance four. This means

$$Y \sim \mathcal{N}(X, 4) \quad \text{with } X \sim \mathcal{U}[0, 5].$$

	optimal	**new1**	**new2**	**CV**
mean (sd) L_1–error	0.103 (0.022)	0.161 (0.040)	0.172 (0.044)	0.194 (0.050)
mean (sd) H	1.520 (0.255)	1.590 (0.658)	1.515 (0.730)	2.750 (0.000)
mean (sd) h	2.195 (0.414)	2.360 (1.010)	2.170 (1.360)	2.590 (1.812)

Table 2: Results of the second example.

We sample again 500 data points. The bandwidths are selected out of sets

$$h \in \mathbf{h} = \{0.5,\ 0.75,\ 1.0,\ 1.25,\ 1.5,\ 1.75,\ 2.0,\ 2.25,\ 2.5,\ 2.75\}$$
$$H \in \mathbf{H} = \{0.5,\ 1.0,\ 1.5,\ 2.0,\ 2.5,\ 3.0,\ 3.5,\ 4.0,\ 4.5,\ 5.0\}.$$

Since the results of our simulation depend on randomly occuring data points, we repeat the whole procedure 100 times. We calculate the estimates of $f_2 : \mathbb{R} \times [1,\ 4] \to \mathbb{R}_+$ and the boxplots in Figure 2 report the average L_1–errors. Table 2 summarises the mean and the standard deviation of the average L_1–errors and chosen bandwidths. In this example the differences are not as noticeable as in the first example. Nevertheless our method

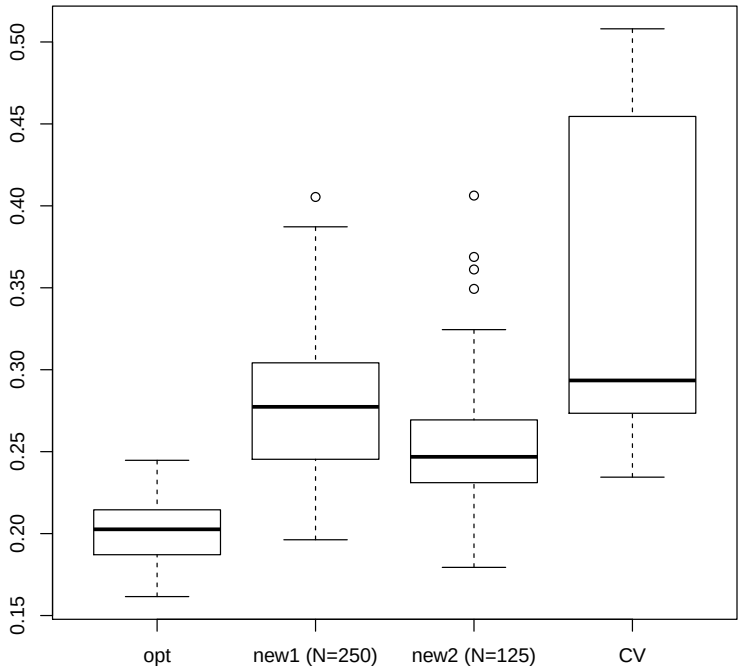

Figure 1: Boxplots of the estimates average L_1–errors in the first example.

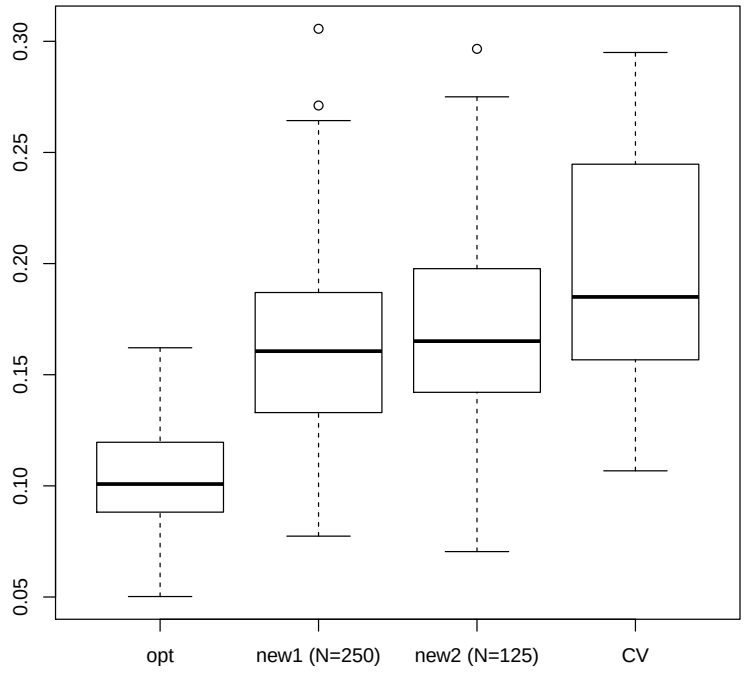

Figure 2: Boxplots of the estimates average L_1–errors in the second example.

shows better results. Again the cross-validation chooses larger mean bandwidths. In case of H, the cross-validation even selects solely the largest bandwidth. In this example the sample size seems to be big enough to avoid noticable differences between the proposed estimates (new1 and new2).

In a third example we sample 200 independent copies of (X, Y), where X is uniformly distributed on $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and Y is chi-squared distributed with degrees of freedom corresponding to the covariate. More precise,

$$Y \sim \chi^2(X) \quad \text{and } X \sim \mathcal{U}\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

In this case the choice of $\mathbf{H}$ is naurally given, since the covariate is discrete on $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. But we evaluate the estimates only on $\{2, 3, 4, 5, 6, 7, 8, 9\}$. The bandwidths are selected out of the finite parameter sets

$$h \in \mathbf{h} = \{0.6,\ 0.7,\ 0.8,\ 0.9,\ 1,\ 1.1,\ 1.2,\ 1.3,\ 1.4,\ 1.5\}$$
$$H \in \mathbf{H} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

As before we compare in Figure 3 the average L_1–errors for 100 repetitions and summarised results are given in Table 3. In this case our proposed estimates outperform the

cross-validation method considerably. The cross-validation selects an remarkble large bandwidth H, while h is smaller in mean compared to the other estimates. The second proposed estimate (new2) seems to have an advantage compared to the first (new1) due to the smaller total sample size.

	optimal	**new1**	**new2**	**CV**
mean (sd) L_1–error	0.218 (0.027)	0.326 (0.046)	0.302 (0.059)	0.563 (0.020)
mean (sd) H	1.600 (0.492)	1.630 (0.720)	1.720 (0.996)	8.000 (0.000)
mean (sd) h	1.473 (0.066)	1.141 (0.275)	1.062 (0.328)	0.976 (0.344)

Table 3: Results of the third example.

To sum up, our estimates outperform the cross-validation method in all three examples. Except of the second example the differences are remarkable. Furthermore, an asymmetric splitting of the sample seems to be advisable if the total sample size is small.

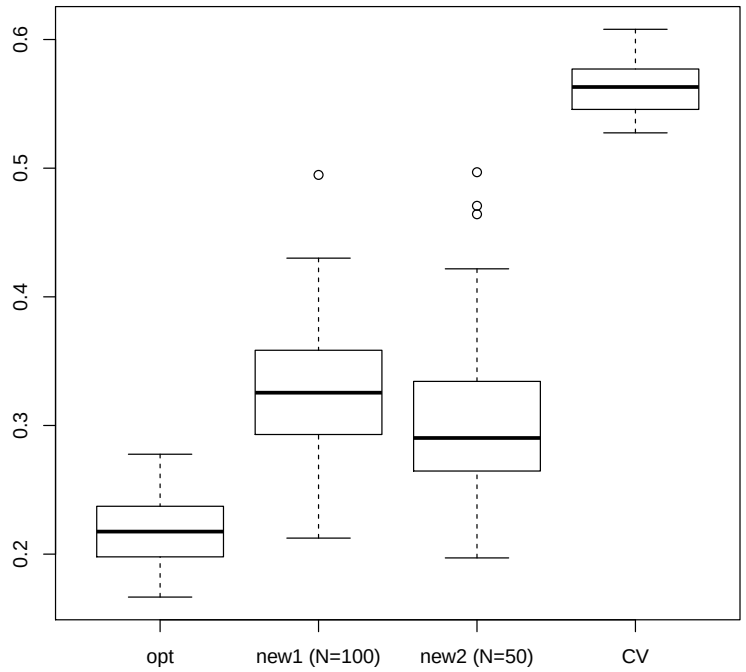

Figure 3: Boxplots of the estimates average L_1–errors in the third example.

4 Proofs

The following lemma is the basis of the proof of Theorem 1 and a generalization of (3) (c.f., Theorem 10.1, Devroye and Lugosi (2001)).

Lemma 1 *It holds for all* $n > 1$

$$\begin{aligned}&\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\hat{\theta}}(y, X_i) - f(y, X_i)| \, dy \\ \leq & 5 \cdot \inf_{\theta \in \mathcal{P}_n} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta}(y, X_i) - f(y, X_i)| \, dy + 4\Delta + \frac{2}{n},\end{aligned}$$

where

$$\Delta = \sup_{\theta_1, \theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int_{A_i(\theta_1, \theta_2)} f(y, X_i) \, dy - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{1}_{A_i(\theta_1, \theta_2)}(Y_i) \right|.$$

Proof. We choose $\theta^* \in \mathcal{P}_n$ arbitrary. With the triangle inequality we get

$$\begin{aligned}&\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\hat{\theta}}(y, X_i) - f(y, X_i)| \, dy \\ \leq & \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\hat{\theta}}(y, X_i) - f_{\theta^*}(y, X_i)| \, dy + \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, X_i) - f(y, X_i)| \, dy \\ =: & A_n + B_n.\end{aligned}$$

By definition $f_{\theta^*}(\cdot, x)$ is a density for all $x \in \mathbb{R}^d$. $\hat{f}_{\hat{\theta}}(\cdot, x)$ is also a density, with the exception of cases where $\hat{f}_{\hat{\theta}}(y, x) = 0$ for all $y \in \mathbb{R}$. In this case

$$\int |\hat{f}_{\hat{\theta}}(y, x) - f_{\theta^*}(y, x)| \, dy = \int |f_{\theta^*}(y, x)| \, dy = \int (f_{\theta^*}(y, x) - \hat{f}_{\hat{\theta}}(y, x))_+ \, dy.$$

If $\hat{f}_{\hat{\theta}}(\cdot, x)$ is a density, we can apply the Lemma of Scheffé and obtain

$$\int |\hat{f}_{\hat{\theta}}(y, x) - f_{\theta^*}(y, x)| \, dy = 2 \cdot \int (f_{\theta^*}(y, x) - \hat{f}_{\hat{\theta}}(y, x))_+ \, dy.$$

Hence,

$$\begin{aligned}A_n = & \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\hat{\theta}}(y, X_i) - f_{\theta^*}(y, X_i)| \, dy \\ \leq & 2 \cdot \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int \left(f_{\theta^*}(y, X_i) - \hat{f}_{\hat{\theta}}(y, X_i) \right)_+ dy\end{aligned}$$

$$
\begin{aligned}
=&2 \cdot \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int_{A_i(\theta^*, \hat{\theta})} f_{\theta^*}(y, X_i)\, dy - 2 \cdot \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int_{A_i(\theta^*, \hat{\theta})} \hat{f}_{\hat{\theta}}(y, X_i)\, dy \\
\leq& 2 \cdot \sup_{\theta_1, \theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int_{A_i(\theta_1, \theta_2)} \hat{f}_{\theta^*}(y, X_i)\, dy - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{1}_{A_i(\theta_1, \theta_2)}(Y_i) \right| \\
&+ 2 \cdot \sup_{\theta_1, \theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int_{A_i(\theta_1, \theta_2)} \hat{f}_{\hat{\theta}}(y, X_i)\, dy - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{1}_{A_i(\theta_1, \theta_2)}(Y_i) \right| \\
\leq& 4 \cdot \sup_{\theta_1, \theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int_{A_i(\theta_1, \theta_2)} \hat{f}_{\theta^*}(y, X_i)\, dy - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{1}_{A_i(\theta_1, \theta_2)}(Y_i) \right| + \frac{2}{n} \\
\leq& 4 \cdot \sup_{\theta_1, \theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int_{A_i(\theta_1, \theta_2)} \hat{f}_{\theta^*}(y, X_i)\, dy - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int_{A_i(\theta_1, \theta_2)} f(y, X_i)\, dy \right| \\
&+ 4 \cdot \sup_{\theta_1, \theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int_{A_i(\theta_1, \theta_2)} f(y, X_i)\, dy - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{1}_{A_i(\theta_1, \theta_2)}(Y_i) \right| + \frac{2}{n} \\
=:& C_n + 4\Delta + \frac{2}{n}
\end{aligned}
$$

According to the Lemma of Scheffé we have

$$
2 \cdot \sup_{B_i \in \mathcal{B}} \left| \int_{B_i} \hat{f}_{\theta^*}(y, x)\, dy - \int_{B_i} f(y, x)\, dy \right| = \int \left| \hat{f}_{\theta^*}(y, x) - f(y, x) \right| dy,
$$

if $\hat{f}_{\theta^*}(\cdot, x)$ is a density. Otherwise $\hat{f}_{\theta^*}(y, x) = 0$ for all $y \in \mathbb{R}$ and we have

$$
\sup_{B_i \in \mathcal{B}} \left| \int_{B_i} \hat{f}_{\theta^*}(y, x)\, dy - \int_{B_i} f(y, x)\, dy \right| = \sup_{B_i \in \mathcal{B}} \left| \int_{B_i} f(y, x)\, dy \right| = \int \left| \hat{f}_{\theta^*}(y, x) - f(y, x) \right| dy.
$$

Thus, it holds

$$
\begin{aligned}
C_n \leq& 4 \cdot \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \sup_{B_i \in \mathcal{B}} \left| \int_{B_i} \hat{f}_{\theta^*}(y, X_i)\, dy - \int_{B_i} f(y, X_i)\, dy \right| \\
\leq& 4 \cdot \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int \left| \hat{f}_{\theta^*}(y, X_i) - f(y, X_i) \right| dy \\
=& 4 \cdot B_n
\end{aligned}
$$

Hence,

$$
\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\hat{\theta}}(y, X_i) - f(y, X_i)|\, dy
$$

$$\leq 5 \cdot \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, X_i) - f(y, X_i)| \, dy + 4\Delta + \frac{2}{n},$$

for all $\theta^* \in \mathcal{P}_n$. □

Proof of Theorem 1. With Lemma 1 we can conclude that

$$\begin{aligned}
&\mathbf{E}\left\{ \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\hat{\theta}}(y, X_i) - f(y, X_i)| \, dy \right\} \\
\leq & 5 \cdot \mathbf{E}\left\{ \inf_{\theta \in \mathcal{P}_n} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta}(y, X_i) - f(y, X_i)| \, dy \right\} + 4 \cdot \mathbf{E}(\Delta) + \frac{2}{n} \\
\leq & 5 \cdot \inf_{\theta \in \mathcal{P}_n} \mathbf{E}\left\{ \int \int |\hat{f}_{\theta}(y, x) - f(y, x)| \, dy \, \mathbf{P}_X(dx) \right\} + 4 \cdot \mathbf{E}(\Delta) + \frac{2}{n}.
\end{aligned}$$

With Lemma 2 below we can conclude that

$$\begin{aligned}
&\mathbf{E}\left\{ \int \int |\hat{f}_{\hat{\theta}}(y, x) - f(y, x)| \, dy \, \mathbf{P}_X(dx) \right\} \\
= & \mathbf{E}\left\{ \int \int |\hat{f}_{\hat{\theta}}(y, x) - f(y, x)| \, dy \, \mathbf{P}_X(dx) - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\hat{\theta}}(y, X_i) - f(y, X_i)| \, dy \right\} \\
& + \mathbf{E}\left\{ \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\hat{\theta}}(y, X_i) - f(y, X_i)| \, dy \right\} \\
\leq & \mathbf{E}\left\{ \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\hat{\theta}}(y, X_i) - f(y, X_i)| \, dy \right\} + 48 \sqrt{\frac{\log n}{\lfloor n/2 \rfloor}} + \frac{242}{\sqrt{\lfloor n/2 \rfloor \cdot \log n}}.
\end{aligned}$$

Due to Lemma 3 below it holds

$$\mathbf{E}(\Delta) \leq 17 \sqrt{\frac{\log n}{\lfloor n/2 \rfloor}} + \frac{16}{\sqrt{\lfloor n/2 \rfloor \cdot \log n}}.$$

Hence, the assertion results from the following two lemmas. □

Remark 4. In the following two lemmas there may be some measureability problems because the supremum is taken over a possible uncountable set. Since in applications only countable sets are considered, we will ignore these problems and refer to van der Vaart and Wellner (1996), where such problems are handled by using the notion of outer probability.

Lemma 2 *It holds for all $n > 1$*

$$\mathbf{E}\left\{ \sup_{\theta \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta}(y, X_i) - f(y, X_i)| \, dy - \int \int |\hat{f}_{\theta}(y, x) - f(y, x)| \, dy \, \mathbf{P}_X(dx) \right| \right\}$$

$$\leq 48\sqrt{\frac{\log n}{\lfloor n/2 \rfloor}} + \frac{242}{\sqrt{\lfloor n/2 \rfloor \cdot \log n}}.$$

Proof. We proceed analogously to the proof of Theorem 9.1 in Györfi et al. (2002).
Step 1: Let $\epsilon > \sqrt{\frac{32}{\lfloor n/2 \rfloor}}$ and $\bar{X}_1, \ldots, \bar{X}_{\lfloor n/2 \rfloor}, X_1, \ldots, X_{\lfloor n/2 \rfloor}$ be independent and identically distributed such that $\bar{X}_1, \ldots, \bar{X}_{\lfloor n/2 \rfloor}$ is independent of $\mathcal{D}_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. We select $\theta^* \in \mathcal{P}_n$ in dependence of $\mathcal{D}_n$ such that

$$\left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, X_i) - f(y, X_i)| \, dy - \int \int |\hat{f}_{\theta^*}(y, x) - f(y, x)| \, dy \, \mathbf{P}_X(dx) \right| > \epsilon.$$

If no $\theta^* \in \mathcal{P}_n$ exists such that the above condition is fulfilled, we choose $\theta^* \in \mathcal{P}_n$ arbitrary. Due to the independence of $\bar{X}_1, \ldots, \bar{X}_{\lfloor n/2 \rfloor}$ it holds

$$\begin{aligned}
&\mathbf{P}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, \bar{X}_i) - f(y, \bar{X}_i)| \, dy - \mathbf{E}\left\{ \int |\hat{f}_{\theta^*}(y, X) - f(y, X)| \, dy \middle| \mathcal{D}_n \right\} \right| > \frac{\epsilon}{2} \middle| \mathcal{D}_n \right\} \\
=&\mathbf{E}\left\{ \mathbb{1}_{\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, \bar{X}_i) - f(y, \bar{X}_i)| \, dy - \mathbf{E}\left\{ \int |\hat{f}_{\theta^*}(y, X) - f(y, X)| \, dy \middle| \mathcal{D}_n \right\} \right| > \frac{\epsilon}{2} \right\}} \middle| \mathcal{D}_n \right\} \\
=&\mathbf{E}\left\{ \mathbb{1}_{\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, \bar{X}_i) - f(y, \bar{X}_i)| \, dy - \mathbf{E}\left\{ \int |\hat{f}_{\theta^*}(y, X) - f(y, X)| \, dy \middle| \mathcal{D}_n \right\} \right|^2 > \frac{\epsilon^2}{4} \right\}} \middle| \mathcal{D}_n \right\} \\
\leq&\mathbf{E}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, \bar{X}_i) - f(y, \bar{X}_i)| \, dy - \mathbf{E}\left\{ \int |\hat{f}_{\theta^*}(y, X) - f(y, X)| \, dy \middle| \mathcal{D}_n \right\} \right|^2 \cdot \frac{4}{\epsilon^2} \middle| \mathcal{D}_n \right\} \\
=&\frac{4}{\epsilon^2 \lfloor n/2 \rfloor^2} \cdot \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbf{E}\left\{ \left| \int |\hat{f}_{\theta^*}(y, \bar{X}_i) - f(y, \bar{X}_i)| \, dy - \mathbf{E}\left\{ \int |\hat{f}_{\theta^*}(y, X) - f(y, X)| \, dy \middle| \mathcal{D}_n \right\} \right|^2 \middle| \mathcal{D}_n \right\} \\
\leq&\frac{4}{\epsilon^2 \lfloor n/2 \rfloor^2} \cdot \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbf{E}\left\{ \left| \int |\hat{f}_{\theta^*}(y, \bar{X}_i) - f(y, \bar{X}_i)| \, dy \right|^2 \middle| \mathcal{D}_n \right\} \\
\leq&\frac{16}{\epsilon^2 \lfloor n/2 \rfloor} \leq \frac{1}{2},
\end{aligned}$$

where we have used that $\int |\hat{f}_{\theta^*}(y, \bar{X}_i) - f(y, \bar{X}_i)| \, dy \leq 2$ for all $i \in \{1, \ldots, \lfloor n/2 \rfloor\}$. Thereby and with the definition of θ^* we have

$$\begin{aligned}
&\mathbf{P}\left\{ \sup_{\theta \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta}(y, X_i) - f(y, X_i)| \, dy - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta}(y, \bar{X}_i) - f(y, \bar{X}_i)| \, dy \right| > \frac{\epsilon}{2} \right\} \\
\geq&\mathbf{P}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, X_i) - f(y, X_i)| \, dy - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, \bar{X}_i) - f(y, \bar{X}_i)| \, dy \right| > \frac{\epsilon}{2} \right\}
\end{aligned}$$

$$
\begin{aligned}
\geq & \mathbf{P}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, X_i) - f(y, X_i)|\, dy - \mathbf{E}\left\{ \int |\hat{f}_{\theta^*}(y, X) - f(y, X)|\, dy \middle| \mathcal{D}_n \right\} \right| > \epsilon, \right. \\
& \left. \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, \bar{X}_i) - f(y, \bar{X}_i)|\, dy - \mathbf{E}\left\{ \int |\hat{f}_{\theta^*}(y, X) - f(y, X)|\, dy \middle| \mathcal{D}_n \right\} \right| \leq \frac{\epsilon}{2} \right\} \\
= & \mathbf{E}\left\{ \mathbb{1}_{\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, X_i) - f(y, X_i)|\, dy - \mathbf{E}\left\{ \int |\hat{f}_{\theta^*}(y, X) - f(y, X)|\, dy \middle| \mathcal{D}_n \right\} \right| > \epsilon \right\}} \cdot \right. \\
& \left. \mathbf{P}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, \bar{X}_i) - f(y, \bar{X}_i)|\, dy - \mathbf{E}\left\{ \int |\hat{f}_{\theta^*}(y, X) - f(y, X)|\, dy \middle| \mathcal{D}_n \right\} \right| \leq \frac{\epsilon}{2} \middle| \mathcal{D}_n \right\} \right\} \\
\geq & \mathbf{E}\left\{ \mathbb{1}_{\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, X_i) - f(y, X_i)|\, dy - \mathbf{E}\left\{ \int |\hat{f}_{\theta^*}(y, X) - f(y, X)|\, dy \middle| \mathcal{D}_n \right\} \right| > \epsilon \right\}} \cdot \frac{1}{2} \right\} \\
= & \frac{1}{2} \cdot \mathbf{P}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta^*}(y, X_i) - f(y, X_i)|\, dy - \mathbf{E}\left\{ \int |\hat{f}_{\theta^*}(y, X) - f(y, X)|\, dy \middle| \mathcal{D}_n \right\} \right| > \epsilon \right\} \\
= & \frac{1}{2} \cdot \mathbf{P}\left\{ \sup_{\theta \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta}(y, X_i) - f(y, X_i)|\, dy - \int \int |\hat{f}_{\theta}(y, x) - f(y, x)|\, dy\, \mathbf{P}_X(dx) \right| > \epsilon \right\}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
& \mathbf{P}\left\{ \sup_{\theta \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta}(y, X_i) - f(y, X_i)|\, dy - \int \int |\hat{f}_{\theta}(y, x) - f(y, x)|\, dy\, \mathbf{P}_X(dx) \right| > \epsilon \right\} \\
\leq & 2 \cdot \mathbf{P}\left\{ \sup_{\theta \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta}(y, X_i) - f(y, X_i)|\, dy - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta}(y, \bar{X}_i) - f(y, \bar{X}_i)|\, dy \right| > \frac{\epsilon}{2} \right\}.
\end{aligned}
$$

Step 2: Let $U_1, \ldots, U_{\lfloor n/2 \rfloor}$ be random variables with

$$
\mathbf{P}\{U_i = 1\} = \mathbf{P}\{U_i = -1\} = \frac{1}{2} \quad (i = 1, \ldots, \lfloor n/2 \rfloor),
$$

such that $U_1, \ldots, U_{\lfloor n/2 \rfloor}, (X_1, Y_1), \ldots, (X_n, Y_n), \bar{X}_1, \ldots, \bar{X}_{\lfloor n/2 \rfloor}$ are independent. Since the joint distribution of of $X_1, \ldots, X_{\lfloor n/2 \rfloor}$ and $\bar{X}_1, \ldots, \bar{X}_{\lfloor n/2 \rfloor}$ is not affected if one randomly interchanges $X_i, \bar{X}_i$, $(i \in \{1, \ldots, \lfloor n/2 \rfloor\})$, we have

$$
\mathbf{P}\left\{ \sup_{\theta \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta}(y, X_i) - f(y, X_i)|\, dy - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\theta}(y, \bar{X}_i) - f(y, \bar{X}_i)|\, dy \right| > \frac{\epsilon}{2} \right\}
$$

$$
\begin{aligned}
=&\mathbf{P}\left\{\sup_{\theta\in\mathcal{P}_n}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}U_i\left(\int|\hat{f}_\theta(y,X_i)-f(y,X_i)|\,dy-\int|\hat{f}_\theta(y,\bar{X}_i)-f(y,\bar{X}_i)|\,dy\right)\right|>\frac{\epsilon}{2}\right\}\\
\leq&\mathbf{P}\left\{\sup_{\theta\in\mathcal{P}_n}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}U_i\cdot\int|\hat{f}_\theta(y,X_i)-f(y,X_i)|\,dy\right|>\frac{\epsilon}{4}\right\}\\
&+\mathbf{P}\left\{\sup_{\theta\in\mathcal{P}_n}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}U_i\cdot\int|\hat{f}_\theta(y,\bar{X}_i)-f(y,\bar{X}_i)|\,dy\right|>\frac{\epsilon}{4}\right\}\\
=&2\cdot\mathbf{P}\left\{\sup_{\theta\in\mathcal{P}_n}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}U_i\cdot\int|\hat{f}_\theta(y,X_i)-f(y,X_i)|\,dy\right|>\frac{\epsilon}{4}\right\}.
\end{aligned}
$$

Step 3: Because of the independence of $U_1,\ldots,U_{\lfloor n/2\rfloor}$ and $\mathcal{D}_n$ we can conclude by the Theorem of Fubini that

$$
\begin{aligned}
&\mathbf{P}\left\{\sup_{\theta\in\mathcal{P}_n}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}U_i\cdot\int|\hat{f}_\theta(y,X_i)-f(y,X_i)|\,dy\right|>\frac{\epsilon}{4}\right\}\\
&=\int\mathbf{P}\left\{\sup_{\theta\in\mathcal{P}_n}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}U_i\cdot\int|\bar{f}_\theta(y,x_i)-f(y,x_i)|\,dy\right|>\frac{\epsilon}{4}\right\}d\bigotimes_{i=1}^{n}P_{(X_i,Y_i)}(x_i,y_i)
\end{aligned}
$$

where for $\theta=(h,H)$

$$
\bar{f}_\theta(y,x)=\frac{\sum_{i=\lfloor n/2\rfloor+1}^{n}K\left(\frac{\|x-x_i\|}{H}\right)K\left(\frac{|y-y_i|}{h}\right)}{h_n\sum_{i=\lfloor n/2\rfloor+1}^{n}K\left(\frac{\|x-x_j\|}{H}\right)}.
$$

Now we want to replace the supremum of an infinite set by a supremum of a finite set. Therefore, we try to find a finite family $\{g_\alpha\}_\alpha$ of functions $g_\alpha:\mathbb{R}\times\mathbb{R}^d\to\mathbb{R}_+$ such that for any $\theta_1=(h_1,H_1)\in\Theta$ we find g_α with

$$
\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\left|\int|\bar{f}_{(h_1,H_1)}(y,x_i)-g_\alpha(y,x_i)|\,dy\right|\leq\frac{\epsilon}{8}.
$$

Let $(x_1,y_1),\ldots,(x_n,y_n)\in\mathbb{R}^d\times\mathbb{R}$ be fixed. At first we let $h_1>0$ be fixed and consider $\bar{f}_{(h_1,H_1)}(y,x_i),i=1,\ldots,\lfloor n/2\rfloor$:

$$
\bar{f}_{(h_1,H_1)}(y,x_i)=\sum_{j=\lfloor n/2\rfloor+1}^{n}W_j(x_i)\cdot\frac{1}{h_1}K\left(\frac{|y-y_j|}{h_1}\right),
$$

with

$$
W_j(x)=\frac{K\left(\frac{\|x-x_j\|}{H_1}\right)}{\sum_{k=\lfloor n/2\rfloor+1}^{n}K\left(\frac{\|x-x_k\|}{H_1}\right)}=\frac{\mathbb{1}_{\{\|x-x_j\|\leq H_1\}}}{\sum_{k=\lfloor n/2\rfloor+1}^{n}\mathbb{1}_{\{\|x-x_k\|\leq H_1\}}}.
$$

The number of different vectors $W_j = \big(W_j(x_1), \dots, W_j(x_{\lfloor n/2 \rfloor})\big)$ for arbitrary $H_1 > 0$ is bounded by the number of matrices of the form

$$\delta = (\delta_{i,j})_{\substack{i=1,\dots,\lfloor n/2 \rfloor \\ j=\lfloor n/2 \rfloor+1,\dots,n}} = \Big(\mathbb{1}_{\{\|x_i - x_j\| \le H_1\}}\Big)_{\substack{i=1,\dots,\lfloor n/2 \rfloor \\ j=\lfloor n/2 \rfloor+1,\dots,n}} \in \{0,1\}^{\lfloor n/2 \rfloor \times \lceil n/2 \rceil}. \tag{6}$$

This follows from the fact that W_j, $(j = \{\lfloor n/2 \rfloor + 1, \dots, n\})$ is a function of δ. For small $H_1 > 0$ all entries are zero, except the entries where $x_i = x_j$ for $i \in \{1, \dots, \lfloor n/2 \rfloor\}$, $j \in \{\lfloor n/2 \rfloor, \dots, n\}$. If one entry $\delta_{i,j}$ changes with increasing H_1, it becomes one and stays one for larger values of H_1. Hence, there are $L_n := \lfloor n/2 \rfloor \cdot \lceil n/2 \rceil + 1$ possibilities for the matrix (6) that we denote by $\delta_{i,j}^{(m)}$, $m = 1, \dots, L_n$. Now we define a class of functions

$$g_{h_1}^{(m)}(y, x_i) = \sum_{j=\lfloor n/2 \rfloor+1}^{n} \frac{\delta_{i,j}^{(m)}}{\sum_{k=\lfloor n/2 \rfloor+1}^{n} \delta_{i,k}^{(m)}} \cdot \frac{1}{h_1} K\left(\frac{|y - y_j|}{h_1}\right)$$

for $i = 1 \dots, \lfloor n/2 \rfloor$, $m = 1, \dots, L_n$ such that for all $i \in \{1 \dots, \lfloor n/2 \rfloor\}$ there exists $m^* \in \{1, \dots, L_n\}$ with

$$\bar{f}_\theta(y, x_i) = g_{h_1}^{(m^*)}(y, x_i).$$

Now we let $H_1 > 0$ be fix and $h_1 \in [1/n, n]$ be arbitrary. Let G_n be an equidistant grid on $[1/n, n]$ with width $\frac{1}{n\sqrt{8n}}$. We choose $h^* = \arg\min_{h \in G_n} |h_1 - h|$ and thus, $|h_1 - h^*| \le \frac{1}{n\sqrt{8n}}$. Since K is the naive kernel it holds

$$\begin{aligned}
&\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int \left| \bar{f}_{(h_1,H_1)}(y, x_i) - \bar{f}_{(h^*,H_1)}(y, x_i) \right| dy \\
=&\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int \left| \sum_{j=\lfloor n/2 \rfloor+1}^{n} W_j(x_i) \cdot \left(\frac{1}{h_1} K\left(\frac{|y - y_j|}{h_1}\right) - \frac{1}{h^*} K\left(\frac{|y - y_j|}{h^*}\right) \right) \right| dy \\
\le&\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \sum_{j=\lfloor n/2 \rfloor+1}^{n} W_j(x_i) \cdot \int \left| \frac{1}{h_1} K\left(\frac{|z|}{h_1}\right) - \frac{1}{h^*} K\left(\frac{|z|}{h^*}\right) \right| dz \\
\le&\int \left| \frac{1}{h_1} K\left(\frac{|z|}{h_1}\right) - \frac{1}{h^*} K\left(\frac{|z|}{h^*}\right) \right| dz \cdot \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \sum_{j=\lfloor n/2 \rfloor+1}^{n} W_j(x_i) \\
\le&\int \left| \frac{1}{h_1} K\left(\frac{|z|}{h_1}\right) - \frac{1}{h^*} K\left(\frac{|z|}{h^*}\right) \right| dz \cdot 1 \\
\le&\int \left| \frac{1}{h_1} K\left(\frac{|z|}{h_1}\right) - \frac{1}{h^*} K\left(\frac{|z|}{h^*}\right) \right| dz + \int \left| \frac{1}{h^*} K\left(\frac{|z|}{h_1}\right) - \frac{1}{h^*} K\left(\frac{|z|}{h^*}\right) \right| dz \\
=&\left| \frac{1}{h_1} - \frac{1}{h^*} \right| \int K\left(\frac{|z|}{h_1}\right) dz + \frac{1}{h^*} \int \left| K\left(\frac{|z|}{h_1}\right) - K\left(\frac{|z|}{h^*}\right) \right| dz \\
=&\left| \frac{1}{h_1} - \frac{1}{h^*} \right| \cdot h_1 + \frac{1}{h^*} \cdot |h_1 - h^*| \\
=&\frac{2}{h^*} \cdot |h_1 - h^*|
\end{aligned}$$

$$\leq 2n \cdot \frac{1}{n\sqrt{8n}} = \frac{2}{\sqrt{8n}} \leq \frac{\epsilon}{8}.$$

Hence, for all $\theta = (h_1, H_1) \in \mathcal{P}_n$ there exist $m \in \{1, \ldots, L_n\}$ and $h \in G_n$ with

$$g_h^{(m)}(y, x_i) = \sum_{j=\lfloor n/2 \rfloor + 1}^{n} \frac{\delta_{i,j}^{(m)}}{\sum_{k=\lfloor n/2 \rfloor + 1}^{n} \delta_{i,k}^{(m)}} \cdot \frac{1}{h} K\left(\frac{|y - y_j|}{h}\right),$$

$i = 1 \ldots, \lfloor n/2 \rfloor$, such that

$$\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left| \int |\bar{f}_{(h_1,H_1)}(y, x_i) - g_h^{(m)}(y, x_i)| \, dy \right| \leq \frac{\epsilon}{8}.$$

By repeated application of the triangle inequality we conclude

$$\begin{aligned}
&\mathbf{P}\left\{ \sup_{\theta \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \int |\bar{f}_\theta(y, x_i) - f(y, x_i)| \, dy \right| > \frac{\epsilon}{4} \right\} \\
\leq &\mathbf{P}\left\{ \sup_{\substack{m \in \{1,\ldots,L_n\} \\ h \in G_n}} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \int |\hat{g}_h^{(m)}(y, x_i) - f(y, x_i)| \, dy \right| + \frac{\epsilon}{8} > \frac{\epsilon}{4} \right\} \\
\leq &\mathbf{P}\left\{ \sup_{\substack{m \in \{1,\ldots,L_n\} \\ h \in G_n}} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \int |\hat{g}_h^{(m)}(y, x_i) - f(y, x_i)| \, dy \right| > \frac{\epsilon}{8} \right\} \\
\leq &L_n \cdot |G_n| \cdot \sup_{\substack{m \in \{1,\ldots,L_n\} \\ h \in G_n}} \mathbf{P}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \int |\hat{g}_h^{(m)}(y, x_i) - f(y, x_i)| \, dy \right| > \frac{\epsilon}{8} \right\} \\
\leq &\sqrt{8} n^{9/2} \cdot \sup_{\substack{m \in \{1,\ldots,L_n\} \\ h \in G_n}} \mathbf{P}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \int |\hat{g}_h^{(m)}(y, x_i) - f(y, x_i)| \, dy \right| > \frac{\epsilon}{8} \right\},
\end{aligned}$$

where the last inequality follows from a trivial computation.
Step 4: The random variables

$$Z_i := U_i \cdot \int |\hat{g}_h^{(m)}(y, x_i) - f(y, x_i)| \, dy \quad (i = 1, \ldots, \lfloor n/2 \rfloor)$$

are independent, take values in $[-2, 2]$ and satisfy $\mathbf{E}\{Z_i\} = 0$ by definition of U_i. With the Hoeffding inequality we can conclude

$$\begin{aligned}
&\mathbf{P}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \int |\hat{g}_h^{(m)}(y, x_i) - f(y, x_i)| \, dy \right| > \frac{\epsilon}{8} \right\} \\
\leq &2 \cdot \exp\left(-\frac{2\lfloor n/2 \rfloor (\epsilon^2/64)}{4^2} \right) = 2 \cdot \exp\left(-\frac{\lfloor n/2 \rfloor \epsilon^2}{512} \right).
\end{aligned}$$

Therfore we have for $\epsilon = \sqrt{\frac{2304 \log n}{\lfloor n/2 \rfloor}}$

$$
\begin{aligned}
&\mathbf{E}\left\{ \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int |\hat{f}_{\hat{\theta}}(y, X_i) - f(y, X_i)|\, dy - \int\int |\hat{f}_{\hat{\theta}}(y, X) - f(y, X)|\, dy\, \mathbf{P}_X(dx) \right\} \\
\leq & \epsilon + 4\sqrt{8} n^{9/2} \int_\epsilon^\infty \mathbf{P}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \int |\hat{g}_h^{(m)}(y, x_i) - f(y, x_i)|\, dy \right| > \frac{u}{8} \right\} du \\
\leq & \epsilon + 4\sqrt{8} n^{9/2} \int_\epsilon^\infty 2 \cdot \exp\left(-\frac{\lfloor n/2 \rfloor u^2}{512} \right) du \\
\leq & \epsilon + 8\sqrt{8} n^{9/2} \int_\epsilon^\infty \exp\left(-\frac{\lfloor n/2 \rfloor \epsilon u}{512} \right) du \\
\leq & \epsilon + \frac{4096\sqrt{8} n^{9/2}}{\lfloor n/2 \rfloor \cdot \epsilon} \exp\left(-\frac{\lfloor n/2 \rfloor \epsilon^2}{512} \right) \\
\leq & 48 \sqrt{\frac{\log n}{\lfloor n/2 \rfloor}} + \frac{242}{\sqrt{\lfloor n/2 \rfloor \cdot \log n}}.
\end{aligned}
$$

□

Lemma 3 *It holds for all* $n > 1$

$$\mathbf{E}(\Delta) \leq 17 \sqrt{\frac{\log n}{\lfloor n/2 \rfloor}} + \frac{16}{\sqrt{\lfloor n/2 \rfloor \cdot \log n}}.$$

Proof. Step 1: Let $\epsilon > \sqrt{\frac{8}{\lfloor n/2 \rfloor}}$. For given $X_1, \ldots, X_{\lfloor n/2 \rfloor}$ we let $Y_1, \ldots, Y_{\lfloor n/2 \rfloor}, \bar{Y}_1, \ldots, \bar{Y}_{\lfloor n/2 \rfloor}$ be conditionally independent and $\bar{Y}_i$ be distributed as Y_i (for given X_i) for all $i \in \{1, \ldots, \lfloor n/2 \rfloor\}$. Now we select $\bar{\theta}_1, \bar{\theta}_2 \in \mathcal{P}_n$ in dependence of $\mathcal{D}_n := \big\{(X_1, Y_1), \ldots, (X_{\lfloor n/2 \rfloor}, Y_{\lfloor n/2 \rfloor})\big\}$ such that

$$\left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int_{A_i(\bar{\theta}_1, \bar{\theta}_2)} f(y, X_i)\, dy - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{1}_{A_i(\bar{\theta}_1, \bar{\theta}_2)}(Y_i) \right| > \epsilon.$$

If no $\bar{\theta}_1, \bar{\theta}_2 \in \mathcal{P}_n$ exists such that the above condition is fulfilled, we choose arbitrary $\bar{\theta}_1, \bar{\theta}_2 \in \mathcal{P}_n$. We define

$$Z_i := \int_{A_i(\bar{\theta}_1, \bar{\theta}_2)} f(y, X_i)\, dy - \mathbb{1}_{A_i(\bar{\theta}_1, \bar{\theta}_2)}(\bar{Y}_i)$$

for all $i \in \{1, \ldots, \lfloor n/2 \rfloor\}$ and thus, $\mathbf{E}\{Z_i | \mathcal{D}_n\} = 0$. Furthermore,

$$\mathbf{P}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} Z_i \right| > \frac{\epsilon}{2} \middle| \mathcal{D}_n \right\}$$

$$
\begin{aligned}
=&\mathbf{P}\left\{\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} Z_i - \mathbf{E}\{Z_i|\,\mathcal{D}_n\}\right| > \frac{\epsilon}{2}\middle|\,\mathcal{D}_n\right\}\\
=&\mathbf{E}\left\{\mathbb{1}_{\left\{\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} Z_i-\mathbf{E}\{Z_i|\,\mathcal{D}_n\}\right|^2>\frac{\epsilon^2}{4}\right\}}\middle|\,\mathcal{D}_n\right\}\\
\leq&\mathbf{E}\left\{\frac{4}{\epsilon^2}\cdot\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} Z_i - \mathbf{E}\{Z_i|\,\mathcal{D}_n\}\right|^2\cdot\mathbb{1}_{\left\{\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} Z_i-\mathbf{E}\{Z_i|\,\mathcal{D}_n\}\right|^2>\frac{\epsilon^2}{4}\right\}}\middle|\,\mathcal{D}_n\right\}\\
\leq&\frac{4}{\epsilon^2}\mathbf{E}\left\{\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} Z_i - \mathbf{E}\{Z_i|\,\mathcal{D}_n\}\right|^2\middle|\,\mathcal{D}_n\right\}\\
=&\frac{V\left\{\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} Z_i\middle|\,\mathcal{D}_n\right\}}{\epsilon^2/4}.
\end{aligned}
$$

Because of the conditional independence of $\bar{Y}_1,\ldots,\bar{Y}_{\lfloor n/2\rfloor}$ it holds

$$
\begin{aligned}
V\left\{\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} Z_i\middle|\,\mathcal{D}_n\right\} =&V\left\{\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\left(\int_{A_i(\bar{\theta}_1,\bar{\theta}_2)} f(y,X_i)\,dy - \mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(\bar{Y}_i)\right)\middle|\,\mathcal{D}_n\right\}\\
=&\frac{1}{\lfloor n/2\rfloor^2}\sum_{i=1}^{\lfloor n/2\rfloor} V\left\{\int_{A_i(\bar{\theta}_1,\bar{\theta}_2)} f(y,X_i)\,dy - \mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(\bar{Y}_i)\middle|\,\mathcal{D}_n\right\}\\
\leq&\frac{1}{\lfloor n/2\rfloor^2}\sum_{i=1}^{\lfloor n/2\rfloor}\mathbf{E}\left\{\left|\int_{A_i(\bar{\theta}_1,\bar{\theta}_2)} f(y,X_i)\,dy - \mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(\bar{Y}_i)\right|^2\middle|\,\mathcal{D}_n\right\}\\
\leq&\frac{1}{\lfloor n/2\rfloor^2}\sum_{i=1}^{\lfloor n/2\rfloor} 1 = \frac{1}{\lfloor n/2\rfloor}.
\end{aligned}
$$

With $\epsilon > \sqrt{\frac{8}{\lfloor n/2\rfloor}}$ we can conclude that

$$
\mathbf{P}\left\{\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\left(\int_{A_i(\bar{\theta}_1,\bar{\theta}_2)} f(y,X_i)\,dy - \mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(\bar{Y}_i)\right)\right| > \frac{\epsilon}{2}\middle|\,\mathcal{D}_n\right\} \leq \frac{4}{\lfloor n/2\rfloor\epsilon^2} \leq \frac{1}{2}.
$$

Thereby and by the definition of $\bar{\theta}_1,\bar{\theta}_2$ we have

$$
\begin{aligned}
&\mathbf{P}\left\{\sup_{\theta_1,\theta_2\in\mathcal{P}_n}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\left(\mathbb{1}_{A_i(\theta_1,\theta_2)}(Y_i) - \mathbb{1}_{A_i(\theta_1,\theta_2)}(\bar{Y}_i)\right)\right| > \frac{\epsilon}{2}\right\}\\
\geq&\mathbf{P}\left\{\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\left(\mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(Y_i) - \mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(\bar{Y}_i)\right)\right| > \frac{\epsilon}{2}\right\}
\end{aligned}
$$

$$
\begin{aligned}
\geq & \mathbf{P}\Bigg\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left(\mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(Y_i) - \int_{A_i(\bar{\theta}_1,\bar{\theta}_2)} f(y, X_i)\, dy \right) \right| > \epsilon, \\
& \qquad \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left(\int_{A_i(\bar{\theta}_1,\bar{\theta}_2)} f(y, X_i)\, dy - \mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(\bar{Y}_i) \right) \right| \leq \frac{\epsilon}{2} \Bigg\} \\
= & \mathbf{E}\Bigg\{ \mathbf{P}\Bigg\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left(\mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(Y_i) - \int_{A_i(\bar{\theta}_1,\bar{\theta}_2)} f(y, X_i)\, dy \right) \right| > \epsilon, \\
& \qquad \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left(\int_{A_i(\bar{\theta}_1,\bar{\theta}_2)} f(y, X_i)\, dy - \mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(\bar{Y}_i) \right) \right| \leq \frac{\epsilon}{2} \,\Bigg|\, \mathcal{D}_n \Bigg\} \Bigg\} \\
= & \mathbf{E}\Bigg\{ \mathbb{1}_{\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left(\mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(Y_i) - \int_{A_i(\bar{\theta}_1,\bar{\theta}_2)} f(y,X_i)\, dy \right) \right| > \epsilon \right\}} \cdot \\
& \qquad \mathbf{P}\Bigg\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left(\int_{A_i(\bar{\theta}_1,\bar{\theta}_2)} f(y, X_i)\, dy - \mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(\bar{Y}_i) \right) \right| \leq \frac{\epsilon}{2} \,\Bigg|\, \mathcal{D}_n \Bigg\} \Bigg\} \\
\geq & \mathbf{E}\Bigg\{ \mathbb{1}_{\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left(\mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(Y_i) - \int_{A_i(\bar{\theta}_1,\bar{\theta}_2)} f(y,X_i)\, dy \right) \right| > \epsilon \right\}} \cdot \frac{1}{2} \Bigg\} \\
= & \frac{1}{2} \cdot \mathbf{P}\Bigg\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left(\mathbb{1}_{A_i(\bar{\theta}_1,\bar{\theta}_2)}(Y_i) - \int_{A_i(\bar{\theta}_1,\bar{\theta}_2)} f(y, X_i)\, dy \right) \right| > \epsilon \Bigg\} \\
= & \frac{1}{2} \cdot \mathbf{P}\Bigg\{ \sup_{\theta_1,\theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left(\mathbb{1}_{A_i(\theta_1,\theta_2)}(Y_i) - \int_{A_i(\theta_1,\theta_2)} f(y, X_i)\, dy \right) \right| > \epsilon \Bigg\}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
& \mathbf{P}\Bigg\{ \sup_{\theta_1,\theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left(\mathbb{1}_{A_i(\theta_1,\theta_2)}(Y_i) - \int_{A_i(\theta_1,\theta_2)} f(y, X_i)\, dy \right) \right| > \epsilon \Bigg\} \\
\leq & 2 \cdot \mathbf{P}\left\{ \sup_{\theta_1,\theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \left(\mathbb{1}_{A_i(\theta_1,\theta_2)}(Y_i) - \mathbb{1}_{A_i(\theta_1,\theta_2)}(\bar{Y}_i) \right) \right| > \frac{\epsilon}{2} \right\}.
\end{aligned}
$$

Step 2: Let $U_1, \ldots, U_{\lfloor n/2 \rfloor}$ be independent random variables with

$$
\mathbf{P}\{U_i = 1\} = \mathbf{P}\{U_i = -1\} = \frac{1}{2}, \quad i = 1, \ldots, \lfloor n/2 \rfloor
$$

such that $U_1, \ldots, U_{\lfloor n/2 \rfloor}$ is independent of $(X_1, Y_1), \ldots, (X_{\lfloor n/2 \rfloor}, Y_{\lfloor n/2 \rfloor})$ and $\bar{Y}_1, \ldots, \bar{Y}_{\lfloor n/2 \rfloor}$. Because of the conditional independence of $Y_1, \ldots, Y_{\lfloor n/2 \rfloor}, \bar{Y}_1, \ldots, \bar{Y}_{\lfloor n/2 \rfloor}$, the conditional

joint distribution of $Y_1, \ldots, Y_{\lfloor n/2 \rfloor}$ and $\bar{Y}_1, \ldots, \bar{Y}_{\lfloor n/2 \rfloor}$ is not affected if one randomly interchanges $Y_i, \bar{Y}_i$, $(i \in \{1, \ldots, \lfloor n/2 \rfloor\})$. Let $X_1^n = \{X_1, \ldots, X_n\}$. Then, it holds

$$
\begin{aligned}
&\mathbf{P}\left\{ \sup_{\theta_1,\theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{1}_{A_i(\theta_1,\theta_2)}(Y_i) - \mathbb{1}_{A_i(\theta_1,\theta_2)}(\bar{Y}_i) \right| > \frac{\epsilon}{2} \right\} \\
=&\mathbf{E}\left\{ \mathbf{P}\left\{ \sup_{\theta_1,\theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{1}_{A_i(\theta_1,\theta_2)}(Y_i) - \mathbb{1}_{A_i(\theta_1,\theta_2)}(\bar{Y}_i) \right| > \frac{\epsilon}{2} \middle| X_1^n \right\} \right\} \\
=&\mathbf{E}\left\{ \mathbf{P}\left\{ \sup_{\theta_1,\theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \left(\mathbb{1}_{A_i(\theta_1,\theta_2)}(Y_i) - \mathbb{1}_{A_i(\theta_1,\theta_2)}(\bar{Y}_i) \right) \right| > \frac{\epsilon}{2} \middle| X_1^n \right\} \right\} \\
\leq&\mathbf{P}\left\{ \sup_{\theta_1,\theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \mathbb{1}_{A_i(\theta_1,\theta_2)}(Y_i) \right| > \frac{\epsilon}{4} \right\} \\
&+ \mathbf{P}\left\{ \sup_{\theta_1,\theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \mathbb{1}_{A_i(\theta_1,\theta_2)}(\bar{Y}_i) \right| > \frac{\epsilon}{4} \right\} \\
=&2 \cdot \mathbf{P}\left\{ \sup_{\theta_1,\theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \mathbb{1}_{A_i(\theta_1,\theta_2)}(Y_i) \right| > \frac{\epsilon}{4} \right\}.
\end{aligned}
$$

Step 3: Because of the independence of $U_1, \ldots, U_{\lfloor n/2 \rfloor}$ and $\mathcal{D}_n$ we can conclude by the Theorem of Fubini that

$$
\begin{aligned}
&\mathbf{P}\left\{ \sup_{\theta_1,\theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \mathbb{1}_{A_i(\theta_1,\theta_2)}(Y_i) \right| > \frac{\epsilon}{4} \right\} \\
=&\int \mathbf{P}\left\{ \sup_{\theta_1,\theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \mathbb{1}_{A_i^*(\theta_1,\theta_2)}(y_i) \right| > \frac{\epsilon}{4} \right\} d \bigotimes_{i=1}^{n} P_{(X_i,Y_i)}(x_i, y_i),
\end{aligned}
$$

where

$$
A_i^*(\theta_1, \theta_2) = \left\{ y \in \mathbb{R} \,:\, \hat{f}_{\theta_1}(y, x_i) > \hat{f}_{\theta_2}(y, x_i) \right\}.
$$

Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ be fixed. Now we count how many different values the vector

$$
\left(\mathbb{1}_{A_1^*(\theta_1,\theta_2)}(y_1), \ldots, \mathbb{1}_{A_{\lfloor n/2 \rfloor}^*(\theta_1,\theta_2)}(y_{\lfloor n/2 \rfloor}) \right) \in \{0,1\}^{\lfloor n/2 \rfloor} \tag{7}
$$

can have for $\theta_1 = (h_1, H_1), \theta_2 = (h_2, H_2) \in \mathcal{P}_n$. The number of different vectors

$$
\left(\mathbb{1}_{\left\{ \hat{f}_{\theta_1}(y_j, x_j) > \hat{f}_{\theta_2}(y_j, x_j) \right\}} \right)_{j=1,\ldots,\lfloor n/2 \rfloor}
$$

with $\theta_1 = (h_1, H_1), \theta_2 = (h_2, H_2) \in \mathcal{P}_n$ is upper bounded by the number of different vectors

$$\left(\mathbb{1}_{\{\bar{f}_{\theta_1,c_1}(y_j,x_j) > \bar{f}_{\theta_2,c_2}(y_j,x_j)\}}\right)_{j=1,\dots,\lfloor n/2 \rfloor}$$

with arbitrary $c_1, c_2 > 0$ and

$$\bar{f}_{\theta_1,c_1}(y_j, x_j) := c_1 \cdot \sum_{i=\lfloor n/2 \rfloor+1}^{n} K\left(\frac{\|x_j - x_i\|}{H_1}\right) K\left(\frac{|y_j - y_i|}{h_1}\right).$$

Now we count the number of different vectors of the form

$$v_{\theta_1} = \left(\sum_{i=\lfloor n/2 \rfloor+1}^{n} K\left(\frac{\|x_1 - x_i\|}{H_1}\right) K\left(\frac{|y_1 - y_i|}{h_1}\right), \dots, \right.$$
$$\left. \sum_{i=\lfloor n/2 \rfloor+1}^{n} K\left(\frac{\|x_{\lfloor n/2 \rfloor} - x_i\|}{H_1}\right) K\left(\frac{|y_{\lfloor n/2 \rfloor} - y_i|}{h_1}\right) \right).$$

This number is upper bounded by the number of matrices of the form

$$(\delta_{i,j})_{\substack{i=1,\dots,\lfloor n/2 \rfloor \\ j=\lfloor n/2 \rfloor+1,\dots,n}} = \left(K\left(\frac{\|x_i - x_j\|}{H_1}\right) K\left(\frac{|y_i - y_j|}{h_1}\right)\right)_{\substack{i=1,\dots,\lfloor n/2 \rfloor \\ j=\lfloor n/2 \rfloor+1,\dots,n}} \in \{0,1\}^{\lfloor n/2 \rfloor \times \lceil n/2 \rceil}.$$

Because if one entry in the vector changes, at least one entry in the matrix must have changed. We now consider the two matrices

$$\delta^{(1)} = (\delta^{(1)}_{i,j})_{\substack{i=1,\dots,\lfloor n/2 \rfloor \\ j=\lfloor n/2 \rfloor+1,\dots,n}} = \left(K\left(\frac{\|x_i - x_j\|}{H_1}\right)\right)_{\substack{i=1,\dots,\lfloor n/2 \rfloor \\ j=\lfloor n/2 \rfloor+1,\dots,n}} \in \{0,1\}^{\lfloor n/2 \rfloor \times \lceil n/2 \rceil}$$

and

$$\delta^{(2)} = (\delta^{(2)}_{i,j})_{\substack{i=1,\dots,\lfloor n/2 \rfloor \\ j=\lfloor n/2 \rfloor+1,\dots,n}} = \left(K\left(\frac{|y_i - y_j|}{h_1}\right)\right)_{\substack{i=1,\dots,\lfloor n/2 \rfloor \\ j=\lfloor n/2 \rfloor+1,\dots,n}} \in \{0,1\}^{\lfloor n/2 \rfloor \times \lceil n/2 \rceil}.$$

It holds

$$\delta_{i,j} = 1 \iff \delta^{(1)}_{i,j} = \delta^{(2)}_{i,j} = 1.$$

For small h_1, H_1 all entries are zero, except the entries where $x_i = x_j$ or $y_i = y_j$ for $i \in \{1, \dots, \lfloor n/2 \rfloor\}$, $j \in \{\lfloor n/2 \rfloor + 1, \dots, n\}$. We now increase h_1, H_1 and count the number of changes. One by another entry turns one until all entries are one. At first we consider $\delta^{(1)}_{i,j}$. With increasing H_1 the entries of $\delta^{(1)}_{i,j}$ become one and stay one. Hence, there are $\lfloor n/2 \rfloor \cdot \lceil n/2 \rceil + 1$ possibilities for this matrix. Analogously, there are $\lfloor n/2 \rfloor \cdot \lceil n/2 \rceil + 1$ possibilities for the matrix $\delta^{(2)}_{i,j}$. Since $\delta_{i,j} = \delta^{(1)}_{i,j} \cdot \delta^{(2)}_{i,j}$, there are at most $(\lfloor n/2 \rfloor \cdot \lceil n/2 \rceil + 1)^2 \leq (n^2/4 + 1)^2 =: L_n$ possibilities for the above matrix $\delta_{i,j}$.
For this reason we now assume that the above vector takes L_n different values $z_1, \dots, z_{L_n} \in$

$\mathbb{R}^{\lfloor n/2 \rfloor}$ for arbitrary $\theta \in \mathcal{P}_n$. There are L_n^2 possibilities to choose $z, \bar{z} \in \{z_1, \dots, z_{L_n}\}$. Let $z = (z^{(1)}, \dots, z^{(\lfloor n/2 \rfloor)}), \bar{z} = (\bar{z}^{(1)}, \dots, \bar{z}^{(\lfloor n/2 \rfloor)})$ and

$$\mathcal{W}_{z,\bar{z}} = \{(\theta_1, \theta_2) \in \mathcal{P}_n \times \mathcal{P}_n | \, (v_{\theta_1}, v_{\theta_2}) = (z, \bar{z})\}.$$

For $(\theta_1, \theta_2) \in \mathcal{W}_{z,\bar{z}}$ and $c_1, c_2 > 0$ it holds:

$$y_i \in \left\{ y \in \mathbb{R} | \, \bar{f}_{\theta_1, c_1}(y, x_i) > \bar{f}_{\theta_2, c_2}(y, x_i) \right\} \iff z^{(i)} > \frac{c_2}{c_1} \bar{z}^{(i)}.$$

Thus, we have for $(\theta_1, \theta_2) \in \mathcal{W}_{z,\bar{z}}$

$$\begin{aligned} &\left\{ \left(\mathbb{1}_{\left\{ \bar{f}_{\theta_1, c_1}(y_j, x_j) > \bar{f}_{\theta_2, c_2}(y_j, x_j) \right\}} \right)_{j=1,\dots,\lfloor n/2 \rfloor} \in \{0,1\}^{\lfloor n/2 \rfloor} \middle| \, c_1, c_2 > 0 \right\} \\ =&\left\{ \left(\mathbb{1}_{\left\{ z^{(1)} > c \cdot \bar{z}^{(1)} \right\}}, \dots, \mathbb{1}_{\left\{ z^{(\lfloor n/2 \rfloor)} > c \cdot z^{(\lfloor n/2 \rfloor)} \right\}} \right) \in \{0,1\}^{\lfloor n/2 \rfloor} \middle| \, c > 0 \right\}. \end{aligned} \tag{8}$$

The number of elements in (8) is upper bounded by $\lfloor n/2 \rfloor + 1$, because for small c the entries of the vector are (nearly) all zero and turn and stay one with growing c. Hence, the number of possible vectors (7) is upper bounded by

$$(\lfloor n/2 \rfloor + 1) \cdot L_n^2 = (\lfloor n/2 \rfloor + 1) \cdot (n^2/4 + 1)^4 \leq n^9 \quad (n > 1).$$

This means the supremum over θ_1, θ_2 is actually the supremum over at most n^9 random variables. Hence,

$$\begin{aligned} &\mathbf{P}\left\{ \sup_{\theta_1, \theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \mathbb{1}_{A_i^*(\theta_1, \theta_2)}(y_i) \right| > \frac{\epsilon}{4} \right\} \\ \leq& n^9 \sup_{\theta_1, \theta_2 \in \mathcal{P}_n} \mathbf{P}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \mathbb{1}_{A_i^*(\theta_1, \theta_2)}(y_i) \right| > \frac{\epsilon}{4} \right\} \end{aligned}$$

Step 4: With the Hoeffding inequality we can conclude

$$\begin{aligned} &\mathbf{P}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \mathbb{1}_{A_i^*(\theta_1, \theta_2)}(y_i) \right| > \frac{\epsilon}{4} \right\} \\ \leq& 2 \cdot \exp\left(- \frac{2 \lfloor n/2 \rfloor (\epsilon^2/16)}{\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} 2^2} \right) = 2 \cdot \exp\left(- \frac{\lfloor n/2 \rfloor \epsilon^2}{32} \right). \end{aligned}$$

Therefore we have

$$\mathbf{P}\left(\Delta > \epsilon \right) \leq 4 \cdot \mathbf{P}\left\{ \sup_{\theta_1, \theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \mathbb{1}_{A_i^*(\theta_1, \theta_2)}(y_i) \right| > \frac{\epsilon}{4} \right\}$$

$$
\begin{aligned}
&\leq 4 \cdot n^9 \sup_{\theta_1,\theta_2 \in \mathcal{P}_n} \mathbf{P}\left\{ \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} U_i \cdot \mathbb{1}_{A_i^*(\theta_1,\theta_2)}(y_i) \right| > \frac{\epsilon}{4} \right\} \\
&\leq 8 \cdot n^9 \exp\left(-\frac{\lfloor n/2 \rfloor \epsilon^2}{32} \right)
\end{aligned}
$$

and with $\epsilon = \sqrt{\frac{288 \log n}{\lfloor n/2 \rfloor}}$ we get

$$
\begin{aligned}
\mathbf{E}(\Delta) &\leq \epsilon + \int_\epsilon^\infty \mathbf{P}(\Delta > u)\, du \\
&\leq \epsilon + \int_\epsilon^\infty 8n^9 \cdot \exp\left(-\frac{\lfloor n/2 \rfloor u^2}{32} \right) du \\
&\leq \epsilon + \int_\epsilon^\infty 8n^9 \cdot \exp\left(-\frac{\lfloor n/2 \rfloor \epsilon u}{32} \right) du \\
&\leq \epsilon + \frac{256\, n^9}{\lfloor n/2 \rfloor \cdot \epsilon} \cdot \exp\left(-\frac{\lfloor n/2 \rfloor \epsilon^2}{32} \right) du \\
&\leq 17 \sqrt{\frac{\log n}{\lfloor n/2 \rfloor}} + \frac{16}{\sqrt{\lfloor n/2 \rfloor \cdot \log n}}.
\end{aligned}
$$

□

References

[1] Bashtannyk, D. M., and Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Computational Statistics and Data Analysis*, **36**, pp. 279–298.

[2] Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, pp. 353–360.

[3] Devroye, L. (1983). The equivalence in L1 of weak, strong and complete convergence of kernel density estimates. *Annals of Statistics*, **11**, pp. 896–904.

[4] Devroye, L. (1987). *A Course in Density Estimation.* Birkhäuser, Basel.

[5] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation. The L1 view.* Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. John Wiley and Sons, New York.

[6] Devroye, L. and Györfi, L. (1990). No empirical probability measure can converge in the total variation sense for all distributions. *Annals of Statistics*, **18**, pp. 1496–1499.

[7] Devroye, L. and Lugosi, G. (1996). A universally acceptable smoothing factor for kernel density estimation. *Annals of Statistics*, **24**, pp. 2499–2512.

[8] Devroye, L. and Lugosi, G. (1997). Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. *Annals of Statistics*, **25**, pp. 2626–2637.

[9] Devroye, L. and Lugosi, G. (1997b). Universal smoothing factor selection in density estimation: theory and practice (with discussion). *Test*, **6**, pp. 223–320.

[10] Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation.* Springer-Verlag, New York.

[11] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* Monographs on Statistics and Applied Probability, Chapman & Hall, London.

[12] Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, **83**, pp. 189–206.

[13] Fan, J. and Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika*, **91**, pp. 819–834.

[14] Gooijer, J. G. D. and Zerom, D. (2003). On conditional density estimation. *Statistica Neerlandica*, **57**, pp. 159–176.

[15] Györfi, L. and Kohler, M. (2007). Nonparametric estimation of conditional distributions. *IEEE Transactions on Information Theory* **53**, pp. 1872–1879.

[16] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer-Verlag, New York.

[17] Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*, **11**, pp. 1156–1174.

[18] Hall, P., Sheater, S. J., Jones, M. C. and Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, **78**, pp. 263–269.

[19] Hall, P., Wolff, R.C.L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, **94**, pp. 154–163.

[20] Holmes, M.P.,Gray, A.G. and Isbell Jr, C.L. (2010). Fast kernel conditional density estimation: A dual-tree Monte Carlo approach. *Computational statistics & data analysis*, **54**, pp. 1707–1718.

[21] Hyndman, R. J., D.M. Bashtannyk and G.K. Grunwald (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, **5**, pp. 315–336.

[22] Hyndman, R.J. and Q. Yao (2002). Nonparametric estimation and symmetry tests for conditional density functions. *Journal of Nonparametric Statistics*, **14**, pp. 259–278.

[23] Mnatsakanov, R. M., and Khmaladze, E. V. (1981). On L_1-convergence of statistical kernel estimators of distribution densities. *Soviet Mathematics Doklady*, **23**, pp. 633–636.

[24] Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, **33**, pp. 1065–1076.

[25] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, pp. 832–837.

[26] Rosenblatt, M. (1969). Conditional probability density and regression estimates. *Multivariate Analysis II* (Ed. P.R. Krishnaiah), Academic Press, New York pp. 25–31.

[27] Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, pp. 65–78.

[28] Scott, D. W. (1992). *Multivariate density estimation: Theory, practice and visualization.* John Wiley, New York.

[29] Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* Monographs on Statistics and Probability, vol. 26. Chapman & Hall, London.

[30] Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, **12**, pp. 1285–1297.

[31] Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics.* Springer-Verlag, New York.